

Sensorveiledning SOS4020

RKH 060318

Årets eksamen er en mappeinnlevering som består av to komponenter – et selvstendig empirisk arbeid (Arbeid 1, 4000 ord) og en artikkel som kommenterer en publisert artikkel (Arbeid 2, 2000 ord). Ordgrenser er +/- 10%, og ordtelling gjøres separat for de to arbeidene. Oppgaveteksten finnes her (og er også vedlagt sensor):

<http://www.uio.no/studier/emner/sv/iss/SOS4020/retningslinjer-for-emneoppgaven.html>

I tillegg til tekst i en PDF-fil leverer studentene log-fil og do-fil (begge for Arbeid 1).

Arbeid 1

Arbeid 1 teller 70%.

Arbeid 1 er et selvstendig empirisk arbeid knyttet til selvvalgt problemstilling. Studentene har levert fem obligatoriske innleveringer som de har fått kommentarer/veiledning på – som så danner grunnlaget for Arbeid 1. Arbeid 1 skal likevel framstå som en helhetlig tekst, ikke som to eller flere «sammenklippede» arbeider. De obligatoriske oppgavene er også vedlagt her. Innholdet i de obligatoriske oppgavene gir en pekepinn på hva som har blitt vektlagt på seminarene.

Studentene har tilgang til tre datasett som alle kan brukes både som tverrsnitt og panel. Datasettene er tilrettelagte versjoner av tre amerikanske paneldatasett (AddHealth, NLSY79 og PSID).

Dokumentasjon for tilrettelegging av datasettene er gjort tilgjengelig for studentene, og vedlagt her.

- PSID har en betydelig mer kompleks struktur enn de andre datasettene, og studentene fikk opprinnelig utlevert en mangelfull versjon av dette datasettet (dette ble rettet opp tidlig i semesteret). PSID inneholder en del muligheter de andre datasettene mangler (linking på tvers av generasjoner og linking av søsken). I vurderingen bør det tas hensyn til at studenter som bruker PSID har hatt en mer krevende jobb med databearbeiding. Den første versjonen av PSID skal ikke benyttes – dette er gitt tydelig beskjed om.
- AddHealth er gjort tilgjengelig i en begrenset versjon (med omfattende dokumentasjon/labels) og en komplett versjon (begrenset dokumentasjon/labels). Studentene har i stor grad brukt kodebøker til AddHealth. Disse er tilgjengelige på nett, men er også gjort tilgjengelige for SOS4020-studenter via Fronter. Studentene skal gjengi informasjon på en slik måte at det ikke skal være nødvendig for sensor å sjekke kodebøker. Kodebøkene legges likevel ved.
- To versjoner av NSLY79 er gjort tilgjengelige for SOS4020: En opprinnelig tilrettelagt versjon (med dokumentasjonsrapport) – denne mangler en variabel for om respondenten har barn. Siden studentene hadde blitt forespeilet en slik variabel, ble det gjort tilgjengelig en enklere versjon av NSLY som inneholder et mål på antall barn.

Studentene har fått beskjed om å kombinere to regresjonsteknikker i oppgaven: En OLS-modell og en logit- eller fasteffektanalyse. Studenter som bruker fasteffekter kan velge å gjennomføre hele oppgaven på paneldatasett (altså en pooled OLS og deretter en fasteffektanalyse), å sammenlikne en OLS fra tverrsnittsdata med fasteffekter på paneldata, eller å sammenlikne en OLS med tverrsnittsdata, pooled OLS med paneldata og fasteffekter med paneldata.

Det er meningen at Arbeid 1 skal være selvforklarende – altså at det ikke skal være nødvendig for sensor å sjekke log- eller do-fil for å forstå hva som er gjort i Arbeid 1. Log-filen fungerer likevel som en ekstra dokumentasjon på hvordan valg mtp. koding og ekskludering av missing etc har fungert. Noen ganger kan pene, håndlagde tabeller skjule underliggende feil i regresjoner, og det kan være nyttig å se logg fra originalkjøringene for å vurdere arbeidet. Do-filen skal sensor kun trenge å bruke unntaksvis, hvis det faktiske er tvil om hvordan resultatene i log-filen kan ha kommet fram.

Det verdsettes at oppgavene

- Stiller et klart forskningsspørsmål som det er mulig å besvare på en god måte med dataene studentene har tilgjengelig.
- Velger adekvate statistiske metoder (innenfor de rammene faget gir), og har god substansiell forståelse av disse.
- Gjør gode valg mtp. avgrensning og bearbeiding av datasettet. Presenterer relevant deskriptiv statistikk.
- Bygger en god regresjonsmodell, med vekt på
 - Et klart skille mellom kontroll- og forklaringsvariabel, der kontrollvariable legges til for å rense ut spuriøsitet i sammenhengen mellom forklaringsvariabel og utfall
 - Motivert inkludering av kontrollvariable, med hensyn til at disse forventes å påvirke både utfall og forklaringsvariabel
 - Et motivert valg av funksjonsform (særlig for forklaringsvariabel, men svært gode oppgaver kan også teste funksjonsform for kontrollvariable) – følgende er gjennomgått: lineær sammenheng, kurvelineær sammenheng, dummypesifikasjon, logaritmisk omkodning av X og Y
- Evner å skille mellom når en beskriver utvalget (ved punkttestimater) og når en generaliserer til populasjonen (ved signifikanstester og/eller konfidensintervaller)
- Adekvat bruk av ulike tester for modellspesifikasjon: Partial og full F-test, LR-test i logistisk regresjon, BIC og AIC.
 - Merk at det feilaktig har blitt oppgitt på forelesning at BIC/AIC kan teste en log-lin og lin-lin modell mot hverandre. Det er ikke tilfelle (og har blitt korrigert), men dette er en feil det bør trekkes lite for. Studentene har begrensede formelle tester for å sammenlikne log-lin og lin-lin, og har også blitt tipset om en «tommefingerregel»: log-lin passer bedre hvis en sammenheng er signifikant i log-lin, men ikke lin-lin, spesifikasjon.
- For studenter som bruker logit-modeller:
 - Kunne tolke **enten** oddsratene, logitter eller marginaleffekter. Gode oppgaver bør kun sammenlikne styrke på marginaleffekter mellom modeller (cf. Mood 2010)

- Kunne bruke Wald-test som et alternativ til t-test, og kjenne til begrensninger ved denne
- Hvis forskningsspørsmålet er kausalt: Har en forståelse av begrensninger ved kontrollvariabelmetoden (og faste effekter hvis dette brukes) for å identifisere effekter. Merk at det ikke kreves for noen karakter at en evner å identifisere effekter – men for gode karakterer er en god diskusjon av hva som evt. hindrer identifikasjon av effekter viktig
- Kunne knytte sammen oppgaven til en helhet, med et motivert forskningsspørsmål som undersøkes empirisk, og en substansiell konklusjon som framhever hva en kan generalisere til populasjonen basert på funnene i oppgaven. En god oppgave må gjerne diskutere forbedringer som kan gjøres i designet som er brukt, og kan også foreslå videre studier.

Arbeid 2

Arbeid 2 teller 30%.

I arbeid 2 skal studentene kommentere en publisert artikkel som bruker en av metodene som dekkes av pensum, men ikke gjennomgås på seminarene. Artikkelen skal være publisert i et tidsskrift på NSD nivå 1 eller 2. Studentene har fått beskjed om å undersøke om tidsskriftet de har valgt er på nivå 1 eller 2 her: <https://dbh.nsd.uib.no/publiseringskanaler/Forside>

På seminarene er det gjennomgått:

- OLS for kontrollvariabelmetoden
- Faste effekter på individ
- Logistisk regresjon

Det vil si at enhver regresjonsteknikk som ikke er en av disse, og som kandidaten kan dokumentere at er omtalt på pensum, kan brukes i Arbeid 2. Sentrale eksempler er (ikke komplett liste):

- Instrumentvariable
- Diff-in-diff
- Regression Discontinuity
- Flernivåmodeller
- Analyse av felteksperimenter, med OLS eller annen regresjonsteknikk

Kandidaten kan velge å kommentere et pensumbidrag, gitt at det tilfredsstillende disse kravene. Hvis kandidaten velger et bidrag utenfor pensum skal dette være tydelig referert, slik at sensor enkelt kan søke det opp.

Dette er en krevende oppgave, og for mange kandidater vil det å vise forståelse av metoden og hvordan den er anvendt være det som er realistisk. For «et godt arbeid» er det også et krav å «drøfte[r] styrker og begrensninger ved både regresjonsteknikken og den konkrete empiriske anvendelsen». Kurset legger i hovedsak opp til substansielle innvendinger («er det dokumentert at pre-reform-trenden i kontroll- og intervensjonsgruppa er lik?»). Innvendinger av mer teknisk art («her er standardfeilene klustret på feil nivå») går utover det som forventes på kurset.

Vedlegg:

- Dokumentasjon av tilrettelagte datasett (6 dokumenter)
- Obligatoriske oppgaver SOS4020 V18 (1 dokument)
- Dokumentasjon av AddHealth (3 dokumenter)
- Retningslinjer og krav for mappeinnlevering (1 dokument)

Sensor kan på oppfordring få tilgang til datasettene som er brukt av SOS4020-studentene.